Cognitive and Achievement Differences Between Students with Divergent Reading and Oral Comprehension Skills: Implications for Accessible Reading Assessment Research

PARA
Partnership for Accessible
Reading Assessment

Cognitive and Achievement Differences Between Students with Divergent Reading and Oral Comprehension Skills: Implications for Accessible Reading Assessment Research

Kevin S. McGrew
Institute for Applied Psychometrics

Ross E. Moen
Martha L. Thurlow
University of Minnesota

May 2010

McGrew, K. S., Moen, R. E., & Thurlow, M. L. (2010). *Cognitive and achievement differences between students with divergent reading and oral comprehension skills: Implications for accessible reading assessment research.* Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

PARA

Partnership for Accessible Reading Assessment
University of Minnesota
207 Pattee Hall
150 Pillsbury Dr. SE
Minneapolis, MN 55455

http://www.readingassessment.info
readingassess@umn.edu

This work is supported, in part, by the U.S. Department of Education, Institute of Education Sciences, National Center for Special Education Research—Grant No. H324F040002. Opinions expressed do not necessarily reflect those of the U.S. Department of Education or offices within it.

ies NATIONAL CENTER FOR SPECIAL EDUCATION RESEARCH
Institute of Education Sciences

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

UNIVERSITY OF MINNESOTA
Curriculum and Instruction

CRESST

UCDAVIS

WESTAT
AN EMPLOYEE-OWNED RESEARCH CORPORATION

The University of Minnesota is committed to the policy that all persons shall have equal access to its programs, facilities, and employment without regard to race, color, creed, religion, national origin, sex, age, marital status, disability, public assistance status, veteran status, or sexual orientation.

This document is available in alternative formats upon request.

Table of Contents

Introduction 1

Models of Reading 1

Measurement of Reading 3

Accessible Reading Assessment Initiatives 4

Purpose of Current Study 5

General Background and Purpose 5

Read-aloud Reading Test Accommodations Research 5

Specific Purpose of Study 8

Reading and Oral Comprehension Measures in WJ III 8

Method 10

Results and Discussion 16

Final CART Tree 16

Variable Importance Rankings 19

Conclusion 20

Summary 20

Study Limitations and Suggestions for Research 21

References 24

Introduction

> The process of text comprehension has always provoked exasperated but nonetheless enthusiastic inquiry within the research community. Comprehension, or "understanding," by its very nature, is a phenomenon that can only be assessed, examined, or observed indirectly…We talk about the "click" of comprehension that propels a reader through a text, yet we never see it directly. We can only rely on indirect symptoms and artifacts of its occurrence. (Pearson & Hamm, 2005, p. 14)

Learning to read is something most all individuals have experienced from their earliest elementary school days. Most all of us use reading skills during a typical day—both at work and at home. Yet, few of us understand the processes involved in extracting meaning from text. For the majority of individuals, reading is an effortless and automatic process. However, for individuals with reading disabilities, reading can be labored and frustrating.

The scientific study of how readers extract meaning from printed text has evolved throughout the 20th century (see Pearson and Hamm, 2005, for a historical overview of reading comprehension theoretical and measurement work). Despite decades of research on the process and products of reading, our understanding and measurement of this ability has proven elusive (Pearson & Hamm, 2005; Shuy, McCardle, & Albro, 2006)

Models of Reading

Contributing to the difficulty in the measurement of reading has been a lack of consensus on the dimensionality of the domain. The scientific study of reading includes a diversity of views about the number and types of component skills and abilities involved during the process (Fletcher, 2006; Shuy et al., 2006). "Because reading is a complex cognitive skill that draws on many component processes and resources, any of these component processes or resources has the potential for being a source of individual differences in reading ability. Some theories of reading ability have emphasized a single component as the major source of individual differences in reading ability. Other theories have emphasized a more multicomponent approach" (Hannon & Daneman, 2001, p. 103).

A review of traditional psychometric factor analytic research has suggested at least five reading subcomponents. In an exhaustive review of the extant factor analytic research of human cognitive abilities, Carroll (1993) identified the following five reading factors:

- *Reading Decoding (RD):* Ability to recognize and decode words or pseudowords in reading using a number of sub-abilities (e.g., grapheme encoding, perceiving multi-letter units, phonemic contrasts, etc.)

- *Reading Comprehension (RC):* Ability to attain meaning (comprehend and understand) connected discourse during reading.

- *Verbal (printed) Language Comprehension (V):* General development, or the understanding of words, sentences, and paragraphs in native language, as measured by reading vocabulary and reading comprehension tests. Does not involve writing, listening to, or understanding spoken information

- *Cloze Ability (CZ):* Ability to read and supply missing words (that have been systematically deleted) from prose passages. Correct answers can only be supplied if the person understands (comprehends) the meaning of the passage

- *Reading Speed (fluency) (RS):* Ability to silently read and comprehend connected text (e.g., a series of short sentences; a passage) rapidly and automatically (with little conscious attention to the mechanics of reading)

Carroll's review focused on literature that addressed primarily "cognitive" or intellectual abilities. Measures of "achievement" (e.g., reading, mathematics, writing) entered his work only to the extent that measures of these constructs were present in the datasets of the cognitive measures. A small number of datasets included, aside from the cognitive variables of interest, variables from domains such as psychomotor performance, perceptual senses (e.g., tactile and olfactory abilities), and a few "conative" (motivational) characteristics. Thus, the five reading factors identified by Carroll are a "bare bones" listing of potentially distinct reading constructs. A larger number of subcomponents would likely be identified in a factor analytic review that focused on a diverse array of reading measures. The lack of consensus on what is "the" list of primary reading components is obvious when one compares Carroll's (1993) reading factor list with the most prominent models of the reading comprehension and decoding processes.

For example, three of the reading factors identified by Carroll (RC, V, CZ) can be classified as subtypes of reading comprehension. Yet, contemporary models of reading comprehension (RC) treat RC as a single unidimensional construct. These RC models range from the popular **Simple View of Reading**, which is two dimensional (RC = reading decoding [D] x listening comprehension [LC]) (Gough, Hoover, & Peterson, 1996; Hoover & Gough, 1990; Neuhaus, Roldan, Boulware-Gooden & Swank, 2006), to a ***Modified Simple View*** (RC = D x LC + speed/fluency [S]) (Joshi & Aaron, 2000), to ***Complex Multidimensional Models*** that, in addition to reading decoding (D) and listening comprehension (LC), hypothesize the involvement of additional skills and abilities (e.g., speed of lexical access or "verbal efficiency"; working memory; cognitive processing speed; vocabulary; prior knowledge; processing capacity; higher-level reasoning processes such as casual inference making, integration, and construction of coherent mental representations of text; memory; and phonological awareness) (Cain, Oakhill & Bryant, 2004; Cain, Oakhill, & Lemmon, 2005; Cutting & Scarborough, 2006; Fletcher, 2006; Francis, Snow, August, Carlson, Miller, & Iglesias, 2006; Hannon & Daneman, 2001; Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003; Neuhaus et al., 2006; RAND Reading Study Group, 2002; Rupp & Lesaux, 2006; van den Broek, Tzeng, Risden, Trabasso, & Basche, 2001). Generally, contemporary accounts of reading comprehension implicitly assume that the comprehension process involves “interpretation of the information in the text, the use of prior knowledge to do so and, ultimately, the construction of a coherent representation or picture of what the text is about in the reader’s mind” (van den Broek, Kendeou, Kremer, Lynch, Butler, White, & Lorch, 2005, p. 109).

Similarly, although Carroll’s (1993) factor analytic review suggested a single reading decoding (RD) factor, competing models of the word identification process exist (e.g., connectionist or “triangle” vs. dual-route models; see Coltheart, 2005; Plaut, 2005). Further supporting a multi-component model of reading is recent heritability research that demonstrates independent genetic influences for RC and RD (Coltheart, 2006) and RD and LC (listening comprehension) (Keenan, Betjemann, Wadsworth, DeFries, & Olson, 2006).

In summary, “reading” is clearly a many splendored thing. There continues to be discussion about a coherent model of reading. At least five reading subconstructs have been identified via factor analysis research. Within these subconstructs, separate “submodels” have been hypothesized and studied (e.g., models of reading comprehension and reading decoding). Even when the smallest set of uniformly mentioned reading subcomponent skills are recognized (e.g., RD and RC), these acknowledgements refer to construct domains where rich and nuanced models (that include additional sub-processes) have been identified. Unfortunately, when it comes to practical reading measures, the specific assessment components that should be used to illustrate this splendor has not yet reached a point of consensus—an often ignored fact in assessment and policy reports. All too often, national and state assessment reports treat reading as if it were a single construct.

Measurement of Reading

Given the diversity of reading models it is not surprising that most contemporary standardized reading tests have an inadequate theoretical basis “which makes them blunt instruments” (Coltheart, 2006, p. 125). The historical twists-and-turns in the development of largely atheoretical reading measures recently resulted in the characterization of reading assessment as “inadequate” (RAND, 2002). For example, the assessment of the reading comprehension subcomponent has been found to vary as a function of type of measurement formats, emphasis (e.g., process vs. products of reading comprehension), differential subcomponent emphasis, and possible instrument-specific inferences (Cutting & Scarborough, 2006; Fletcher, 2006; RAND, 2002; Shuy et al., 2006). Current reading comprehension assessments have been negatively characterized as failing to: (a) adequately represent the complexity of the reading comprehension domain, (b) control for construct irrelevant variance, (c) incorporate developmental process information, (d) incorporate information regarding interests and values, (e) provide instructional assistance to teachers, (f) broaden and richen the academic curriculum, (g) incorporate multidimensional components, and (h) often meet minimal psychometric criteria (reliability and validity) (Fletcher, 2006; RAND, 2002; Shuy et al., 2006).

The heterogeneity of reading models and measures makes it difficult to draw broad inferences from the extant reading research. For example, within the sub-domain of reading comprehension, the RAND study group concluded that "understanding the nature of the problem of reading comprehension requires having available good data identifying which readers can successfully undertake which activities in which texts. Such data are not available, in part because the widely used comprehension assessments are inadequate" (p. 52). Although excellent recommendations exist for improving the state-of-the-art of reading comprehension research and development (e.g., see RAND, 2002), reading comprehension research must continue within the known constraints of both reading comprehension theory and measurement.

Given the reading model and measurement context we have described, we hope that this report will make a small contribution to understanding the process of reading and its measurement. As with most reading research, the current study has known a priori constraint-driven limitations and flaws. Nonetheless, we believe the results of the current investigation, which is driven by the contemporary practical and policy-driven forces to develop *accessible reading assessment for all students*, can contribute to the larger reading measurement puzzle.

Accessible Reading Assessment Initiatives

The current wave of accountability-driven education reform, spurred by the 2001 Elementary and Secondary Education Act (ESEA) (known as the *No Child Left Behind [NCLB] Act*), has created intense interest in assessment supports that will increase the participation of students with disabilities in state assessment programs. Because schools are accountable for demonstrating the reading proficiency of an increasingly diverse population of students, it is important that state group assessments be accessible and accurately measure each student's reading proficiency.

The U.S. Department of Education's Office of Special Education Programs (OSEP) established the *National Accessible Reading Assessment Projects* (NARAP) to conduct research intended to make large-scale assessments of reading proficiency more *accessible* and accurate for students with disabilities. "The goal of these projects [later funded through the Institute of Education Sciences] is to produce research findings and assessment techniques that demonstrate how large-scale assessments of reading proficiency can become more accessible and valid for all students, while also meeting the assessment requirements of …NCLB." A diverse array of NARAP research development activities has occurred. The purpose of the current study is to focus a different, complimentary, lens on a portion of the NARAP research and development activities.

Purpose of Current Study

General Background and Purpose

Anecdotal classroom teacher reports suggest that educators believe that state level reading scores often underestimate the "true" reading proficiency (typically measured as reading comprehension) for some students or the state assessments fail to reflect the progress these students have made in the acquisition of important reading comprehension related skills (e.g., critical evaluation of text). These anecdotal reports are borne out, to some extent, by recent research of the Partnership for Accessible Reading Assessment project (Moen, Liu, Thurlow, Lekwa, Scullin, & Hausmann, 2009) and the New England Compact Enhanced Assessment Initiative (New England Compact, 2007). New England Compact researchers found that they could describe the achievement gaps of students by combining teacher judgments with a state administered assessment. Moen et al. asked teachers to identify the characteristics of students that impeded their reading test performance. They conducted brief assessments and structured interviews with students whose teachers had identified them as being inaccurately measured by state assessments. Results indicated that many students were accurately identified by their teachers, while other students definitely were not. Results indicated that supplemental evidence could be found that supported some, although certainly not all, teachers' judgments about students having reading abilities that would be missed with typical reading tests. Although the sample was small, this preliminary study indicated that there is more than anecdotal evidence that state level reading scores may be underestimating the "true" reading proficiency of some students.

Given the emerging nature of evidence from teacher reports, empirical research that could shed light (even partial light) on the characteristics of students perceived to be "less accurately measured readers (LAMR)" or the characteristics of the assessment tools that contribute to the "he/she-can-read-higher-than-the-state-test-score-says" phenomena (hereafter referred to as the "*LAMR effect*") is potentially important. Not only could such information help identify the types of assessment supports these students may need during state reading tests, this information could also influence the development and revision of state reading assessments.

One ideal investigation of the LAMR effect would be to objectively identify students with disabilities whose performance on state reading tests do not reflect their "true" level of reading proficiency. The cognitive and non-cognitive characteristics of these LAMR students could then be measured and described. In addition, the characteristics of the tests that tend to produce the most notable LAMR-effect could also be studied in this pool of identified students.

Read-aloud Reading Test Accommodations Research

In the current wave of state accountability assessment systems, a variety of test accommodations have been implemented across states for test taking by students with disabilities (Christensen, Lazarus, Crone, & Thurlow, 2008; Thurlow, Lazarus, Thompson, & Morse, 2005). According to these recent state accommodations policy analyses, the most frequently mentioned test accommodation policies can be classified into five broad categories (presentation, response, equipment/materials, scheduling/timing, and setting). One of the more controversial *presentation accommodations,* an accommodation that, interestingly, is also being suggested as a viable *accessible reading* assessment format given that it has been proposed as one form of alternative literacy (Cunningham, 2000), is the provision of "read-aloud" accommodations to students with disabilities *during* <u>*reading*</u> *tests* (Johnstone, Thurlow, Thompson, & Clapper, 2008). In simple terms, this change in procedures removes the demand for a student to decode the text because the text would be "read aloud" to the student and the student would then answer the subsequent comprehension questions. A more complex description of the impact of this read aloud accommodation that takes into account the complex multidimensional models of reading described above could delve into various visual processing demands that are removed by the accommodation and various auditory cues and demands that might be introduced by it.

The idea of using oral comprehension as an accommodation and format for accessible reading tests is grounded in a lengthy history of reading research that has considered listening comprehension (linguistic or language comprehension) as a reliable and valid proxy or predictor of a students reading "ability" or "potential," particularly with increasing amounts of education (Aaron, 1997; Joshi & Aaron, 2000). In simple terms, listening comprehension can be defined as "the ability to understand and relate spoken language to one's personal experience. Listening comprehension relies upon the ability to encode information at a rate that supplies enough factual details so that inferences can be made, and to self-monitor understanding of the oral passage" (Neuhaus et al., 2006, p. 42).

This line of reasoning and research has been the basis for comparisons (often via discrepancy formulas) between a student's listening and reading comprehension. The rationale for this comparison is that an individual cannot understand what he or she is reading unless he or she understands the material when it is read aloud to him or her. According to Aaron (1997, p. 467), "listening comprehension places an upper limit on reading comprehension…the correlation between the two forms of comprehension is high, usually in the vicinity of .80." Although the use of listening comprehension as a proxy for reading comprehension potential or ability is a relatively old concept, the concept is alive and well in contemporary cognitive psychology conceptualizations of reading comprehension. For example, van den Broek, Kendeou, Kremer, Lynch, Butler, White, and Lorch (2005) suggested that the comprehension processes involved in non-text/print and text/print-based comprehension are very similar, even at the preschool age level. These researchers argued for the importance of measuring basic reading comprehension processes via a variety of nontextual materials.

Research has sought to investigate the viability of the reading test read-aloud accommodation for students with disabilities. Huynh and Barton (2006) examined the effect of oral administration accommodations on the internal test factor structure of Grade 10 student performance on the South Carolina High School Exit Examination (HSEE). These researchers reported that internal (factor) structure of the HSEE test was consistent across standard and read-aloud administration formats. In addition, Huynh and Barton (2006) reported that read-aloud HSEE performance of students with disabilities was equal to performance under standard administration conditions. Huynh and Barton (2006) concluded that the read-aloud accommodation "leveled the playing field" for students with disabilities and was a viable reading test accommodation. In contrast, Cook, Eignor, Sawaki, Steinberg, and Cline (2006) reported a mixed set of positive and negative findings (in terms of psychometric integrity for this accommodation) across the five studies of the audio, oral, or read-aloud accommodations in the literature they reviewed.

Adding further to the mixed psychometric landscape of the validity of the reading test read-aloud accommodation is the recent study by Cahalan-Laitusis, Cook, Cline, King, and Sabatini (2008). In relatively large samples of 4th (n = 1181) and 8th (n = 847) grade students in the state of New Jersey, students with and without reading-based learning disabilities took both a standard administration and a read-aloud administration of a reading comprehension test. Cahalan-Laitusis et al. (2008) reported that the mean score on the read-aloud (audio) version was higher than scores on the standard version for both students with and without a reading learning disability across grades 4 and 8. In addition, students with reading disabilities differentially benefited more than students without disabilities at both grades. When examining the impact of the read-aloud accommodation on differential item functioning (DIF) in 4th and 8th grade samples, a subset of the same ETS research group (Pitoniak, Cook, Cline, & Cahalan Laitusis, in press) reported minimal DIF as a function of disability status, a finding supporting the validity of the read-aloud reading test accommodation.

In summary, the feasibility of using read-aloud accommodations on reading comprehension tests is actively being pursued by a variety of researchers and vis-à-vis a variety of research methods. This research reflects an improvement in the state-of-the art of test accommodation research reported by Thurlow and Bolt (2001) when they concluded that "research has primarily supported the use of the read aloud accommodation for students with disabilities on math tests. However, great concern has been expressed about the validity of using this accommodation on reading tests, and limited research has addressed this issue . . . Clearly, more research needs to be done on the oral reading accommodation to determine how it affects what the test measures" (p. 32).

Despite the recent increase in research focused on the psychometric integrity of the reading test read-aloud accommodation, the relatively mixed findings in this area of inquiry suggest the field has yet to reach a consensus on the psychometric viability of this accommodation, and more importantly, the use of this test format as the *accessible reading* method for students with low reading performance or specific reading disabilities in large-scale testing programs. Echoing the conclusion of Thurlow and Bolt (2001) and Thurlow et al. (2005), additional research is needed before the read-aloud accessible reading format bandwagon gathers too much steam.

Specific Purpose of Study

A missing piece of the read-aloud/accessible reading research seems to be information about which students may or may not benefit from this approach. That is, which students, and more specifically, what characteristics of specific students, interact (either positively or negatively) with the read-aloud/accessible format? Stated differently, which students who experience difficulty with reading may or may not benefit from a read-aloud/accessible assessment? Just as many different profiles of knowledge and skills can lead to standards-based reading comprehension proficiency level classifications (Rupp & Lesaux, 2006), it is important to determine whether students, with different profiles, may differentially benefit from read-aloud/accessible reading approaches. The current exploratory study was designed to provide preliminary insights into these questions.

As a co-author of the Woodcock-Johnson III Battery (WJ III; Woodcock, McGrew & Mather, 2001), the first author of this report has access to the large, nationally representative WJ III norm data. The WJ III norm sample spans preschool through late adulthood and includes a diverse array of individually administered cognitive and achievement tests. The question asked in the current investigation was, *"are there any analysis of the WJ III data that might shed light on learner characteristics that differentiate students whose measured reading performance is below what might be considered their optimal/predicted reading performance?"* The research described here addresses this question.

The current investigation operated under the constraints of the available measures in the WJ III norm data. Although the data set is not ideally designed for studying LAMR and MAMR (Less and More Accurately Measured Readers) effects, the results of an analysis of the WJ III data are potentially informative for research and development efforts focused on large-scale accessible reading assessment programs.

Reading and Oral Comprehension Measures in WJ III

The WJ III includes two individually administered tests that share a common testing format but tap the separate and related reading and oral comprehension skills of test takers. The two tests are Passage Comprehension and Oral Comprehension.

Passage Comprehension. This test is designed to measure reading comprehension vis-à-vis a test taker's skill in reading a short passage and identifying a missing keyword. In this *modified cloze* procedure, the subject must exercise a variety of comprehension and vocabulary skills. The test items require examinees to read short passages and to identify a missing key word in the passage. This task requires the examinee to state a word that would be appropriate in the context of the passage. This is an example of a modified cloze procedure.

The *cloze approach* can be defined as "a method of systematically deleting words from a prose selection and then evaluating the success a reader has in accurately supplying the words deleted" (McKenna & Robinson, 1980). The assumption underlying the cloze procedure is that a reader can only supply the correct word if he or she understands (i.e., comprehends) the meaning of the text (Joshi, 1995). The cloze procedure is an attempt to "assess reading comprehension by providing longer, hence more 'real world,' reading experiences during assessment than is possible with other formats" (Osterlind, 1989).

Although a form of the cloze technique was used as early as 1897 to investigate memory, Taylor (1953) is credited with first developing and applying this procedure to the measurement of language or reading proficiency (Carroll, 1993; McKenna & Robinson, 1980; Pearson & Hamm, 2005). Variants of the cloze procedure have been used for a variety of applications (e.g., foreign language; determining readability of text; a teaching device) (McKenna & Robinson, 1980) and was used extensively in reading comprehension research in the 1960s (Pearson & Hamm, 2005). The popularity of cloze techniques has not come without criticism. According to Pearson and Hamm (2005), "the unsettled question about cloze tests is whether they are measures of individual differences in comprehension or measures of the linguistic predictability of the passages to which they are applied. Cloze techniques have been widely criticized for this ambiguity" (p. 24).

Readers interested in in-depth information about the cloze approach should consult McKenna and Robinson's (1980) annotated bibliography that covers (a) background, (b) literature reviews, (c) comprehension and readability, (d) statistical and constructional issues, (e) the psychology of cloze, (f) contextual phenomena, (g) use as a teaching device, (h), foreign language applications, and (i) the cloze and maze procedure. More recent overviews can be found in McGrew (1999) and Pearson and Hamm (2005)

Oral Comprehension. The WJ III Oral Comprehension test measures the ability to listen to a short tape-recorded passage and to verbally supply the single word missing at the end of the passage. It is identical in format and rationale to the previously described Passage Comprehension test, with the primary distinction being that the subject listens to the passages rather than reading them independently. No decoding or fluency skills are involved in the Oral Comprehension test.

The presence of two tests with identical formats, save for the critical distinction of one being presented orally and the other requiring the subject to read passages, provides a unique opportunity to investigate possible learner characteristics that differentiate individuals who display discrepant performance between the two measures.

Why is such a comparison important? In the current wave of state accountability assessment systems, the idea of providing read-aloud accommodations to students with reading problems has been a topic of significant debate. In simple terms, this change in procedures would remove the demand for a student to decode the text, because the text would be "read aloud" to the student and the student would then answer the subsequent comprehension questions. This describes the essential difference between the WJ III Oral Comprehension (similar to a "read aloud" test accommodation) and Passage Comprehension (test taker required to decode) tests.

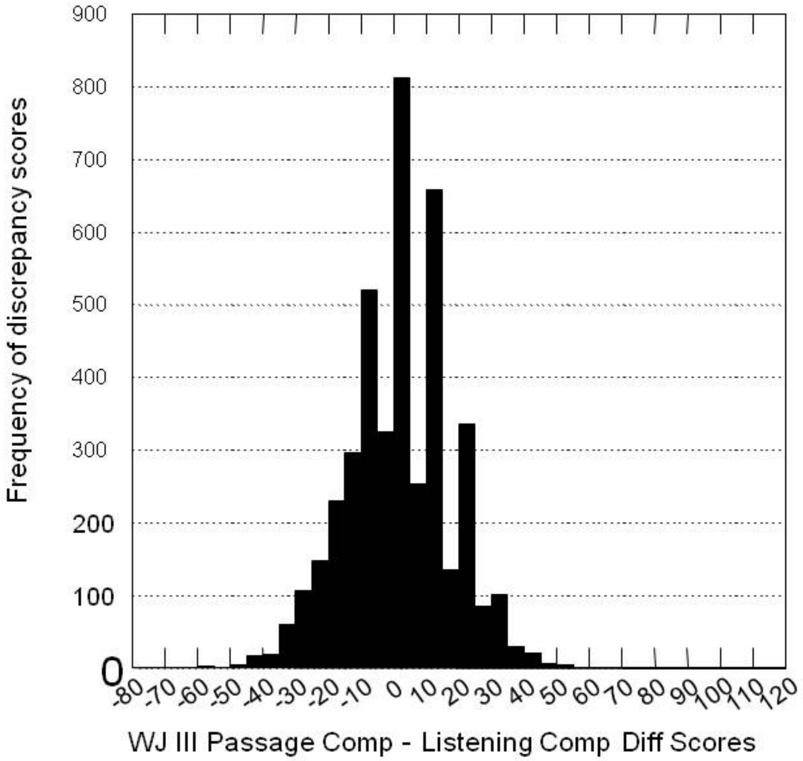
To be clear, the WJ III Passage Comprehension test does not mirror most reading comprehension tests on state large-scale assessments (where a student reads paragraphs and then answers inferential comprehension questions). Even so, there is evidence that the WJ modified cloze Passage Comprehension test has equal validity as a measure of reading comprehension when compared to more traditional reading comprehension tests (see McGrew, 1999). The co-occurrence of a passage comprehension measure with a parallel oral comprehension measure in a large nationally representative database provides the opportunity for a potentially informative exploratory investigation.

Method

Sample. All school-age (grades K-12) individuals in the WJ III norming who had scores for both the Passage Comprehension and Oral Comprehension tests were selected. The WJ III nationally representative standardization sample was constructed using a stratified sampling plan that controlled for 10 individual variables (e.g., race, sex, educational level, occupational status) and community variables (e.g., community size, community socio-economic status) as described by the United States Census projections for the year 2000 (McGrew & Woodcock, 2001). The development, standardization, and psychometric properties of these test batteries have generally been evaluated favorably by independent reviewers (Bradley-Johnson, Morgan, & Nutkins, 2004; Cizek, 2003).

A simple difference score was calculated between the two measures (*Passage Comprehension (PC) Standard Score – Oral Comprehension (OC) Standard Score*). All age-based standard scores (SS) had a mean of 100 and an SD of 15. The simple difference equation was employed so that negative PC-OC difference scores (PCOCDIF) reflected reading comprehension *below* oral comprehension, which operationally defined (in this investigation) a pool of less accurately measured readers (LAMR). A total of 4,177 PCOCDIF scores were obtained from the WJ III K-12 norm sample. PCOCDIF scores ranged from -79 to +120 (M = 1.1; SD = 15.9). Figure 1 displays a frequency histogram of the PCOCDIF scores for the entire sample.

Figure 1. Frequency Histogram of WJ III PCOCDIF Scores Across Grades K-12



Two different groups were operationally defined based on the WJ III PCOCDIF distribution of scores. The first group represented WJ III norm subjects where the PCOCDIF discrepancy between oral comprehension and reading comprehension was greater than or equal to 0 (i.e., zero and all positive scores). *More Accurately Measured Readers* (MAMR) were those whose reading and oral comprehension scores were similar or whose reading comprehension score surpassed the oral comprehension score. Table 1 presents a frequency breakdown of students by group and grade in the WJ III K-12 norm sample. A total of 3,472 subjects were available for analyses.

Using the PCOCDIF distribution standard deviation as a guide (SD of approximately 15), the LAMR group was operationally defined as subjects who displayed reading comprehension (WJ III Passage Comprehension) below oral comprehension (WJ III Oral Comprehension) of at least 2/3 of standard deviation of the distribution of PCOCDIF scores. Thus, LAMR subjects were operationally defined as subjects with PCOCDIF scores less than or equal to -10 points. Subjects with PCOCDIF discrepancies between -9 and -1 were eliminated from the analyses in an attempt to define two distinctly different groups (LAMR vs. MAMR).

Table 1. Frequency Breakdown of Number of WJ III K-12 Norm Subjects by LAMR/MAMR Classification

Grade	PCOCDIF >=0 (MAMR)	PCOCDIF <= -10 (LAMR)	Total N
K	128	99	227
1	147	101	248
2	168	80	248
3	201	108	309
4	253	112	365
5	246	135	381
6	195	95	290
7	211	56	267
8	194	62	256
9	190	45	235
10	194	44	238
11	172	46	218
12	145	45	190
Total	**2444**	**1028**	**3472**

Potential Group Differentiating Variables. Grade placement (GP), in tenths of an academic year, and all of the WJ III cluster or test measures are presented in Table 2. These data were included in exploratory analyses as potential variables that might differentiate LAMR and MAMR students.

Data Analytic Method. Given the absence of a priori empirical or theoretically based hypotheses regarding LAMR/MAMR group differences, the exploratory "data mining" procedure of *classification and regression tree analyses* (CART; Berk, 2009; Brieman, Friedman, Olshen, & Stone, 1984; Sonquist, 1970) was used to explore potential differences between the operationally defined LAMR and MAMR groups.

Data mining is an umbrella term that describes a number of sophisticated computer-intensive statistical non-parametric procedures designed to identify unknown patterns and relationships in large databases. *Predictive data mining* models are typically used to forecast explicit values, based on patterns determined from known results. In contrast, *descriptive data mining* models are used to describe patterns in existing data, and are generally used to identify and describe meaningful subgroups. The descriptive CART methodology was used in this study.

Table 2. Description of WJ III measures used as predictor (independent variables) in LAMR/MAMR CART analysis

Ability domain/WJ III measure	Description of abilities measured
Other reading subskills	
Letter-Word Identification test (LWID)	Single test measure of the ability to read isolated letters and words.
Word Attack test (WA)	Single test measure of the reading ability to apply phonic and structural analysis skills to the pronunciation of unfamiliar printed words.
Reading Fluency test (RF)	Single timed test measure of the reading ability to quickly comprehend the correctness of simple sentences.
Reading Vocabulary test (RV)	Single test measure of the reading ability to understand the meanings of words (antonyms, synonyms & analogies).
General intelligence	
General Intellectual Ability-Standard cluster (GIAS)	Seven-test g-weighted measure of general intelligence based on the seven WJ III Standard cognitive tests.
Language, verbal abilities and general knowledge	
Comprehension-Knowledge cluster (Gc)	Two-test cluster measure of the ability to use language and acquired knowledge effectively.
Knowledge cluster (KN)	Two-test cluster measure of general information and cultural knowledge.
Oral Expression-Standard cluster (OE)	Two-test cluster of linguistic competency and expressive vocabulary ability.
Academic Knowledge test (AK)	Single test measure of knowledge in various areas of the biological and physical sciences, history, geography, government, economics, art, music, and literature.
Story Recall test (SR)	Single test measure of ability to recall increasingly complex orally presented stories presented.
Picture Vocabulary test (PV)	Single test measure of the ability name familiar and unfamiliar pictured objects (vocabulary).
Reasoning	
Fluid Reasoning cluster (Gf)	Two-test cluster measure of the ability to form and recognize logical relationships among patterns, to make deductive and inductive inferences, and to transform novel stimuli.
Numerical Reasoning cluster (NR)	Two-test cluster measure of the ability to reason with mathematical concepts involving the relationships and properties of numbers.

Table 2. Description of WJ III measures used as predictor (independent variables) in LAMR/MAMR CART analysis (continued)

Ability domain/WJ III measure	Description of abilities measured
Visual-spatial abilities/processing	
Visual-Spatial Processing cluster (Gv)	Two-test cluster measure of the ability to recognize spatial relationships and to understand, analyze, store, retrieve, manipulate, and think with stimuli that are presented visually.
Visualization cluster (VIS)	Two-test cluster measure of the ability to envision objects or patterns in space by perceiving how the object would appear if presented in an altered form.
Auditory abilities/processing	
Auditory Processing cluster (Ga)	Two-test cluster measure of the ability to perceive, attend to, and analyze patterns of sound and speech that may be presented in distorted conditions.
Phonemic Awareness cluster (PA)	Two-test cluster measure of the ability to perceive separate units of speech sounds in order to analyze and synthesize those units.
Sound Discrimination cluster (SD)	Two-test cluster measure of the ability to distinguish between pairs of voice-like or musical sound patterns.
Long-term storage and retrieval	
Long-term Retrieval cluster (Glr)	Two-test cluster measure of the ability to store and readily retrieve information in long-term memory.
Associative Memory cluster (AM)	Two-test cluster measure of the ability store and retrieve associations (paired-associate learning).
Short-term and working memory	
Short-term Memory cluster (Gsm)	Two-test cluster measure of the ability to understand and store information in immediate awareness and then use it within a few seconds.
Working Memory cluster (WM)	Two-test cluster measure of the ability to temporarily store and mentally manipulate information held in immediate memory.
Understanding Directions test (UD)	Single test measure of the ability to comprehend linguistic concepts (receptive language) presented via oral directions.
Auditory Memory Span cluster (AMS)	Two-test cluster measure of the ability to listen to and then immediately recall sequentially ordered information after one presentation.
Cognitive processing speed	
Processing Speed cluster (Gs)	Two-test cluster measure of the ability to perform simple cognitive tasks quickly, especially when under pressure to maintain focused attention and concentration.
Perceptual Speed cluster (PS)	Two-test cluster measure of the ability to rapidly scan and compare visual symbols.

The CART® data mining software (v5.0; http://www.salford-systems.com) was applied to the current research study. CART® is a robust decision-tree tool that automatically sifts, via complex iterative mathematical sorting and splitting algorithms, large complex databases, searching for and isolating significant patterns and relationships. The discovered knowledge, if accurate and demonstrating strong cross-validation via *n*-fold internal cross-validation methods, can be used to generate reliable, easy-to-grasp predictive decision-tree models for practical application.

Ma (2005) recently demonstrated the usefulness of CART methods when applied to the analysis of math achievement growth in school-age students. Ma's (2005) brief CART description, as applied to the identification of different math achievement student subgroups, was:

> CART performs binary splitting of groups successively based on a statistical criterion. Starting from the entire sample (called the root node), each explanatory variable is examined for how well it splits students into two groups (called child nodes). CART provides a measure called impurity to guide the splitting. Impurity measures the degree to which students in a node vary in outcome measure. A smaller impurity indicates a more homogeneous outcome for a node. A reduction in impurity can be calculated by comparing impurity of the root node with the sum of impurities of its child nodes. The explanatory variable that yields the largest reduction in impurity is selected for performing the first split. The resulting child nodes are markedly different in outcome measure. Each node is again split through the same procedure (nodes that descend child nodes are called parent nodes). As the process continues, students are classified into smaller and smaller nodes. Similarity in outcome measure within each node increases and, meanwhile, difference in outcome measure between nodes also increases. Nodes that cannot be split further are called terminal nodes. A rule or standard that regulates the meaning of a reasonable reduction in impurity is used for discontinuing the splitting process. When the reduction in impurity becomes smaller than the rule, the parent node is not split and is declared a terminal node. If the rule stops the splitting too soon, then the resulting CART is too small to discover the relationships in the data. If the rule stops the splitting too late, then-the resulting CART is too large to have meanings (having terminal nodes with few students in each). To deal with that problem, CART grows a very large tree first and then prunes the tree by combining nodes on the basis of the reduction in impurity. (p.80)

CART provides distinct advantages over traditional parametric multivariate statistical procedures when employed in an exploratory study using a large database. First, if significant differences exist between the LAMR and MAMR groups, the computing-intensive iterative algorithms will likely discover and describe (via "if-then" decision rules) well-replicated (10-fold internal cross-validation was employed in the current study) and different subgroups. Second, *interactions* between variables can be detected, an important feature when a priori empirical literature is meager. Third, variables that are highly correlated can be included in the same analyses. Problems with multicolinearity and singularity (in the underlying covariance/correlation matrices) in parametric multivariate statistical analyses typically require the elimination of highly correlated variables (especially composite variables that include common measures/variables). Highly correlated and overlapping variables can all be included in a CART analyses. Fourth, in addition to identifying the most important variables that differentiate the primary target groups, different numbers of relatively homogeneous subgroups, if they exist in the data, will be identified. In CART the focus is not on identifying the linear combination(s) of variables that best discriminate between primary target groups (e.g., as in discriminant function analyses), but on finding as many possible homogeneous subgroups that are differentiated by the most discriminating pragmatic "if-then" combination of variables. Finally, variables that are not employed as the "splitting" variables at each successive node, but which are consistently related to group differences, are identified via a pragmatic "variable importance" metric. Briefly, *variable importance* rankings reflect a variable's contribution that stems from both the variable's role as a primary splitter and its role as a surrogate to primary splitters. Conceptually this is akin to tracking each variable's partial correlation at each step in step-wise multiple regression and "counting/ranking" the "almost-a-significant-predictor-at-a-step" status of each variable across all successive steps during model building. Thus, variables that may be important, but that do not emerge as primary splitters, are not lost via a singular focus on the final model.

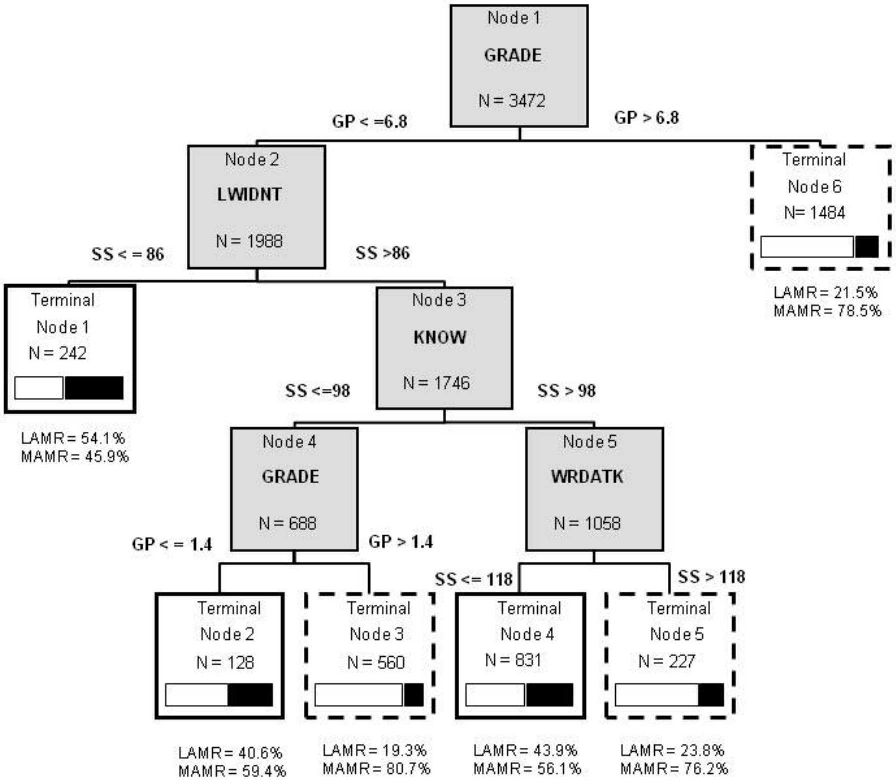
Given the exploratory nature of the current investigation, the desire to include highly correlated (and overlapping) WJ III cognitive and achievement composite measures in a single analysis (e.g., each student's general intelligence score as well as the components that contribute to the IQ score), plus the goal to detect unknown patterns and relationships in a large database, CART methods seemed particularly relevant to the current study.

Results and Discussion

Final CART Tree

Figure 2 presents the final CART decision tree that classified the complete student pool (n=3,472) as either LAMR or MAMR. In Figure 2 the grey box nodes represent the different decision or splitting points based on the variable name listed inside the box (e.g., Node 1 was split based on GRADE—student grade placement). In the final model only four of the original pool of 26 potential variables aided in the classification of operationally defined LAMR/MAMR students. These four variables were (see Table 2 for definitions and abilities measured):

Figure 2. Final CART LAMR/MAMR Decision Tree



- GRADE – Students grade placement in tenths of an academic year
- LWIDNT – WJ III Letter-Word Identification standard score
- KNOW – WJ III Knowledge cluster standard score
- WRDATK – WJ III Word Attack cluster standard score

As summarized in Figure 2, the complete sample of 3,472 subjects was classified into six different subgroups (*terminal nodes*). The *dashed-line white* terminal node boxes (terminal nodes # 3, 5, 6) represent subjects with a predicted MAMR classification. The *solid-line white* terminal node boxes (terminal nodes # 1, 2, 4) represent subjects where the primary classification is LAMR. The *white/black* shaded horizontal rectangular bars (within each terminal node) represent the degree of classification accuracy within the specific node. In a highly discriminating and powerful CART decision tree, the ideal horizontal bars (within each terminal node) would be predominately one color (which would indicate a relatively homogenous subgroup). The percent of each subgroup correctly classified is listed under each terminal node in Figure 2.

The most salient conclusion from an inspection of Figure 2 is that the classification of MAMRs is much more accurate than LAMRs. For example, in the MAMR terminal node subgroups (dashed-line white boxes) the percent of accurate MAMR classification ranged from 76.2 % (terminal node #5) to 80.7 % (terminal node #3). In contrast, LAMR subgroup classification was closer to chance levels (50%; classification accuracy) and ranged from 40.6% (terminal node #2) to 54.1% (terminal node #1). Overall, the CART "prediction success" values for the final model were 73.2% for MAMRs and 46.7% for LAMRs.

This level of prediction success indicates that this optimal decision tree model presented in Figure 2 operates at, or slightly below, chance levels in the identification of the primary target group—LAMR students. This level of accuracy suggests that despite the application of optimal exploratory data mining software to a comprehensive set of cognitive and achievement variables (in a large nationally representative sample), accurate classification of LAMR students was not highly successful. Although limitations to the current study design mitigate broad generalizations to hypothesized LAMR effects in state large scale reading assessments, the current findings suggest that research investigating the LAMR-effect may need to look beyond inherent cognitive and achievement characteristics of students with disabilities who are believed to be inaccurately measured by state reading assessments.

Specific CART-Tree Interpretation. Despite the chance-level classification of LAMR students, interpretation of the CART decision-tree (see Figure 2) provides information that may inform future research. The following general conclusions are drawn from an inspection of Figure 2.

- A subject's grade placement (as operationally defined in the current study) is the single most powerful variable differentiating LAMR and MAMR students. The initial splitting rule of grade placement = 6.8 (see Node 1 in Figure 2) suggests that below or equal to grade 6.8, a greater proportion of school-age students display reading comprehension skills that are lower than their oral comprehension skills. The presence of terminal node #6 (and no subsequent splitting below this node) suggests that after the end of approximately sixth grade, discrepancies between oral and reading comprehension (in favor of the former) are less likely to be found in the general population. Furthermore, after grade 6.8, it is not possible, given the collection of WJ III cognitive and achievement variables employed in this study, to isolate different subgroups of students who display consistent oral and reading comprehension proficiency or reading comprehension skills beyond their oral comprehension skills.

  The finding that grade placement is the most important LAMR/MAMR discriminating variable most likely reflects a development-by-construct interaction effect. It is known that oral comprehension skill development precedes reading comprehension development. In the current model, the initial splitting node at grade 6.8 suggests that this may be the grade (age) level where oral and reading comprehension skills become consistent for most individuals in the general population. In other words, it is not unusual, in the general student population, for reading comprehension to be below oral comprehension for students who are less than or equal to grade 6.8. The single terminal node #6 for subjects beyond grade 6.8 indicates that LAMR/MAMR students cannot be differentiated by the set of cognitive and achievement predictor (IV's) variables used in this analysis. *Other variables (e.g., non-cognitive behaviors and aptitudes; instructional and environmental factors) would need to be examined to determine whether distinct LAMR or MAMR subgroups can be identified beyond grade 6.8.*

- Low sight recognition reading skills (WJ III Letter-Word Identification test standard score less than or equal to 86—approximately1 SD below the mean) increases the probability that a student will display discrepant oral and reading comprehension abilities (LAMR in particular or Node #3). *Low word recognition skills*, although not allowing for precise prediction and description of individual students as LAMR/MAMR, appears to be the variable that needs inclusion in future LAMR-effect research.

    - For students below grade 6.8 with adequate word recognition skills (WJ III Letter-Word Identification standard score > 86), those with low general verbal information (WJ III Knowledge standard score less than average—<98) may be at greater risk for being classified as LAMR, but only during the early school years (grade placement ≤ 1.4—see terminal node #2). However, above grade 1.4, low general knowledge does not appear to be significantly associated with a probable LAMR classification (see terminal node 3). Thus, lower than average general knowledge appears to increase discrepancies between reading comprehension and oral comprehension only at the earliest grade levels. A significant caveat for this interpretation is addressed in the section on "Variable Importance Rankings."
    - Students below grade 6.8 who have adequate word recognition skills (WJ III Letter-Word Identification standard score > 86—Node #2) and average or above general verbal information (WJ III Knowledge standard score > 98—Node #3), in general, are not frequently classified as LAMR, unless they also display a relative deficit in "sounding out" words (WJ III Word Attack standard score ≤118—Node #5). This finding is not a strong finding as the classification accuracy for such students is only at approximately chance levels (43.9%; see terminal node #4). Still, this finding suggests that word attack ability is a variable that needs further exploration in LAMR-effect research.

Variable Importance Rankings

Inspection of the final CART-based variable importance rankings sheds additional light on possible important ability constructs to consider in the design of future LAMR-effect research. The following were the eight most "valuable" potential predictor variables in the current CART analyses.

○ **Grade placement (GP)**	100.0
○ **Knowledge (KNOW)**	46.6
○ Comprehension-Knowledge (Gc)	41.2
○ **Letter-Word Identification (LWID)**	40.3
○ **Word Attack (WA)**	39.4
○ Academic Knowledge (AK)	27.8
○ Oral Expression (OE)	22.5
○ Picture Vocabulary (PV)	20.5

It is important to note that within any CART analysis, the top predictor is assigned a value of 100 and all other variables are scaled relative to the value of 100. The variables in bold designate the variables that were included in the final CART decision tree (see Figure 2)

The most important finding from the above variable importance rankings is that the abilities measured by the four variables that were *not* included in the final CART tree are highly related and within the same general human ability domain (i.e., measures of language, verbal abilities and general knowledge—see Table 2) as the Knowledge (KNOW) predictor. This suggests that if the Knowledge cluster had been omitted from the analyses, one of the other language/verbal/knowledge variables (Gc, AK, OE, PV) would likely served the same function in the model (in CART terminology these other variables would be considered good “surrogate” variables for the primary splitting variable at a node). Additionally, the finding that 5 of the 8 most “important” variables discovered via CART data mining are from the same general human ability domain suggests that a student’s general language development, verbal abilities, and general knowledge may be an important domain to include in future LAMR-effect research.

Conclusion

Summary

Today’s schools are accountable for demonstrating the reading proficiency of an increasingly diverse population of students. In this context, research and development activities are underway to design more *accessible* large scale reading tests for students with disabilities.

Anecdotal classroom teacher reports and emerging research evidence suggest that many educators believe that state level reading scores often underestimate the “true” reading proficiency (typically measured as reading comprehension) for some students with disabilities. Given that these reports are largely anecdotal, the current exploratory study was conducted to shed empirical light on the characteristics of students that educators believe are “*less accurately measured readers (LAMR*).”

Classification and regression tree analysis (CART), a robust exploratory data mining procedure, was used to determine whether significant patterns of cognitive and achievement characteristics (measured by the WJ III battery) could be identified that accurately differentiate and describe LAMR (less accurately measured readers) and MAMR (more accurately measured readers) in a nationally representative sample of K-12 students (WJ III battery standardization sample). The LAMR and MAMR groups were operationally defined by the presence or absence of a significant discrepancy between student scores on similarly formatted individually administered tests of oral comprehension (OC) and reading passage comprehension (PC). This OC-PC operational definition is consistent with the notions that measures of oral or listening comprehension are (a) optimal estimates of a student's ability or potential for reading, and (b) "read aloud" accommodations (for reading tests) should be considered as one means for securing reading proficiency scores for students who have difficulty displaying their "true" reading comprehension on group reading tests.

Despite the application of the powerful exploratory CART methodology to a comprehensive collection of intellectual and achievement ability measures, accurate classification of students as LAMR was not highly successful (at approximately chance levels—40-50% accurate). The findings suggest that research investigating the LAMR-effect may need to look beyond the inherent cognitive and achievement characteristics of students with disabilities who are believed to be inaccurately measured by state reading assessments. Other variables (e.g., non-cognitive behaviors and aptitudes, instructional and environmental factors, large scale test and format characteristics) need to be examined to better understand the variables that may result in some students with disabilities being inaccurately measured in reading. Despite the inability to develop an accurate predictive model (for individual student level prediction), the current study did suggest basic word recognition skills (word sight vocabulary and word attack skills) and general verbal ability and fund of verbal knowledge are characteristics that warrant inclusion in future research directed at understanding the LAMR-effect in group reading tests.

Study Limitations and Suggestions for Research

Several limitations in the current study were noted. For each of them, suggestions for additional research are identified. First, the current study was not conducted on samples of identified students with disabilities. The extent to which the current results, which are based on a nationally representative sample of K-12 students, generalize to specific groups of students with disabilities is unknown and warrants additional study. Second, students were operationally defined as either more accurately (MAMR) or less accurately measured readers (LAMR) based on a discrepancy between two test scores (oral and reading comprehension). The degree to which the LAMR/MAMR operational definition would identify the same students as might be nominated by teachers, or by some other empirically-based identification procedure, is unknown. Third, the measures used to operationally define more accurately and less accurately measured readers were *individually administered* tests. Current accessible reading assessment research and development efforts are focused on large scale *group administered* measures of reading. The degree to which the current findings would generalize to similarly defined groups of students based on group administered oral and reading comprehension tests is unknown and warrants future study.

On the other hand, using individual rather than group administered tests might be viewed as a design strength in the current study instead of a limitation. Given that the individually administered test situation is more tailored to the characteristics of the student being tested, and that the clinical nature of 1-1 testing allows for the elicitation of maximal student test performance, greater confidence can be placed in the conclusions about the relevance of the cognitive and achievement characteristics (to the LAMR-effect) investigated in the current study. If a similar study was conducted with large scale assessments of achievement, the potential confounding influence of construct irrelevant variance on the cognitive and achievement variables, largely due to non-cognitive variables (e.g., attention, motivation, interest, etc.) operating in a large group testing situation would be unknown. In a sense, the current investigation can be considered a more tightly controlled design that reduced the potential confounds of the effects of non-cognitive learner characteristics that likely influence performance during large scale testing for students with disabilities. This suggests that greater rather than less confidence can be placed in the conclusion that cognitive and achievement characteristics may not be the major reasons why some students fail to be measured accurately by typical reading tests.

This points to a fourth limitation of this study; it focused almost exclusively on cognitive and achievement characteristics. Other potentially important test-performance-related student characteristics were missing. The failure to identify robust relations between the cognitive and achievement characteristics of students and more accurately versus less accurately measured reading performance (with the exception of mild effects for word recognition and verbal knowledge abilities), begs for additional research that would include measures of: (a) student background characteristics, (b) important sensory-motor characteristics, (c) general characteristics of the group testing environment that may interact with student characteristics, (d) item format and design characteristics of group tests that may interact with different student characteristics, and (e) non-cognitive learner characteristics that may be important for school learning and optimal test performance.

A study by Thurlow, Moen, Lekwa, and Scullin (2010) supports the conclusion that relying on cognitive characteristics to reduce the LAMR effect may be less fruitful than some might hope. This study was focused on reducing the impact of poor decoding skills on total reading performance. Students used "reading pens" to have words in a reading test pronounced for them. This attempt to provide students with what was labeled a "partial auditory accommodation" failed to improve performance on the reading test. This finding raises questions as to what the sources of improvement in test performance are when more help is provided than strictly decoding assistance.

A study cited earlier (Moen et al., 2009) in which teachers were asked to identify the characteristics of students that impeded their reading test performance suggests that although classic cognitive characteristics may well play a role for a small number of less accurately measured readers, other kinds of variables seem likely to play a larger role for a larger number of students. Teachers described factors such as motivation, engagement, attention, anxiety, and learning styles. A recent research synthesis of *essential student academic facilitators* (McGrew, Johnson, Cosio, & Evans, 2004), as well as other recent research syntheses (e.g., Elliot & Dweck, 2005) should be reviewed for more information about potentially important conative or non-cognitive abilities (e.g., academic motivation, academic goal setting and orientation, reading self-efficacy and self-concept, self-regulated learning/cognitive strategies) that might provide insights into the characteristics of students who experience difficulty accurately performing on large scale reading assessments.

References

Aaron, P. G. (1997). The impending demise of the discrepancy formula. *Review of Educational Research, 67*, 462–502.

Berk, R. (2009). *Statistical learning from a regression perspective*. Springer: New York.

Breiman, L., Friedman, J. H., Olshen, R.A. & Stone, C. J. (1984). *Classification and regression trees.* Pacific Grove, CA: Wadsworth.

Bradley-Johnson, S., Morgan, S. K., & Nutkins, C. (2004). Test review. [Review of The Woodcock–Johnson III Tests of Achievement.] *Journal of Psychoeducational Assessment, 22,* 261–274.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96*(1), 31-42.

Cain, K., Oakhill, J., & Lemmon, K. (2005). The relation between children's reading comprehension level and their comprehension of idioms. *Journal of Experimental Child Psychology, 90*(1), 65-87.

Cahalan-Laitusis, C., Cook, L., Cline, F., King, T., & Sabatini, J. (2008). *Examining the impact of audio presentation on tests of reading comprehension* (ETS RR-08-23). Princeton, NJ: Educational Testing Service.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press.

Christensen, L. L., Lazarus, S. S., Crone, M., & Thurlow, M. L. (2008). *2007 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 69). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Cizek, G. J. (2003). [Review of the Woodcock-Johnson III]. In B. S. Plake & J. C. Impara (Eds.), *The fifteenth mental measurements yearbook* (pp. 1020–1024). Lincoln, NE: Buros Institute of Mental Measurements.

Coltheart, M. (2005). Modeling reading: The dual-route approach. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 6-23). MA, USA: Blackwell Publishing.

Coltheart, M. (2006). The genetics of learning to read. *Journal of Research in Reading, 29*(1), 124-132.

Cook, L., Eignor, D., Sawaki, Y., Steinberg, J., & Cline, F. (2006). *Using factor analysis to investigate the impact of accommodations on the scores of students with disabilities on English-language arts assessments.* Princeton, NJ: Educational Testing Service, Designing Accessible Reading Assessments Project.

Cunningham, J. W. (2000). How will literacy be defined in the new millennium? *Reading Research Quarterly, 35(1),* 64-71.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*(3), 277-299.

Elliot, A. J., & Dweck, C. S. (2005). *Handbook of competence and motivation.* New York: Guilford.

Fletcher, J. M. (2006). Measuring reading comprehension. *Scientific Studies of Reading, 10*(3), 323-330.

Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2006). *Scientific Studies of Reading, 10(3),* 301-322.

Gough, P., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 1-13). Mahwah, NJ: Erlbaum Assoc.

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology, 93*(1), 103-128.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2,* 127-160.

Huynh, H., & Barton, K. E. (2006). Performance of students with disabilities under regular and oral administrations of a high-stakes reading examination. *Applied Measurement in Education, 19*(1), 21-39.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and fluency. *Journal of Educational Psychology, 95*(4), 719-729.

Johnstone, C. J., Thurlow, M. L., Thompson, S. J., & Clapper, A. T. (2008). The potential for multi-modal approaches to reading for students with disabilities as found in state reading standards. *Journal of Disability Policy Studies, 18*(4), 219-229.

Joshi, R. M. (1995). Assessing reading and spelling skills. *School Psychology Review, 24*(3), 361-375.

Joshi, R. M., & Aaron, P. G. (2000). The component model of reading: Simpleview of reading made a little more complex. *Reading Psychology, 21*, 85–97.

Keenan, J. M., Betjemann, R. S., Wadsworth, S. J., DeFries, J. C., & Olson, R. K. (2006). Genetic and environmental influences on reading and listening comprehension. *Journal of Research in Reading, 29*(1), 75-91.

Ma, X. (2005). Growth in mathematics achievement: Analysis with classification and regression trees. *Journal of Educational Research, 99*(2), 78-86.

McKenna, M. C., & Robinson (1980). *An introduction to the cloze procedure: An annotated bibliography*. Newark, DE: International Reading Association.

McGrew, K. (1999). *The measurement of reading achievement by different individually administered standardized reading tests: Apples and apples, or apples and oranges?* (Research Report # 1). St. Cloud, MN: Institute for Applied Psychometrics.

McGrew, K. M., Johnson, D., Cosio, A., & Evans, J. J. (2004). *Increasing the chance of no child being left behind: Beyond cognitive and achievement abilities*. Minneapolis, MN: Institute on Community Integration, University of Minnesota.

McGrew, K. S., & Woodcock, R. W. (2001). *Technical Manual. Woodcock-Johnson III.* Itasca, IL: Riverside Publishing.

Moen, R., Liu, K., Thurlow, M., Lekwa, A., Scullin, S., & Hausmann, K. (2009, May). Identifying less accurately measured students. *Journal of Applied Testing Technology, 10*(2).

Neuhaus, G. F., Roldan, L. W., Boulware-Gooden, R., & Swank, P. R. (2006). Parsimonious reading models: Identifying teachable subskills. *Scientific Studies of Reading, 10(3),* 221-224.

New England Compact. (2007). *Reaching students in the gaps: A study of assessment gaps, students, and alternatives* (Grant CFDA #84.368 from U.S. Department of Education Office of Elementary and Secondary Education to the Rhode Island Department of Education). Newton, MA: Education Development Center.

Osterlind, S. J. (1989). *Constructing test items*. Boston: Kluwer Academic Publishers.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices—Past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13–69). Mahwah, NJ: Lawrence Erlbaum Associates.

Pitoniak, M., Cook, L., Cline, F., & Cahalan Laitusis, C. (in press) *Using differential item functioning to investigate the impact of accommodations on the scores of students with disabilities on English-language arts assessments.* Princeton, NJ: Educational Testing Service.

Plaut, D. C. (2005). Connectivist approaches to reading. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 24-38). Malden, MA: Blackwell Publishing.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R & D program in reading comprehension*. Santa Monica, CA: RAND.

Rupp, A. A., & Lesaux, N. K. (2006). Meeting expectations? An empirical investigation of a standards-based assessment of reading comprehension. *Educational Evaluation and Policy Analysis, 28*(4), 315-333.

Shuy, T. R., McCardle, P., & Albro, E. (2006). Introduction to this special issue: Reading comprehension assessment. *Scientific Studies of Reading, 10(3),* 221-224.

Sonquist, J. A. (1970). *Multivariate model building: The validation of a search strategy.* Ann Arbor, MI: University of Michigan, Institute for Social Research.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 415-433.

Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy* (Synthesis Report 41). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Lazarus, S. S., Thompson, S. J., & Morse, A. B. (2005). State policies on assessment participation and accommodations for students with disabilities. *Journal of Special Education, 38*(4), 232-240.

Thurlow, M. L., Moen, R. E., Lekwa, A. J., & Scullin, S. B. (2010). *Examination of a reading pen as a partial auditory accommodation for reading assessment.* Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

van den Broek, P., Kendeou, P., Kremer, K., Lynch, J., Butler, J., White, M. J., & Lorch, E. P. (2005). In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 107-129). Mahwah, NJ: Lawrence Erlbaum Associates.

van den Broek, P., Tzeng, Y., Risden, K., Trabasso, T., & Basche, P. (2001). Inferential questioning: Effects on comprehension of narrative texts as a function of grade and timing. *Journal of Educational Psychology, 93*(3), 521-529.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Examiner's manual. *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside Publishing.